A RISK-SENSITIVE APPROACH TO POPULATION ETHICS

WESLEY STONE

§ INTRODUCTION

How many people should there ever be? Derek Parfit poses "this awesome question" in *Reasons and Persons* (1984).[1] It seems strange to try to put a precise number on it, yet the question has long held political salience. Traditionally right-wing "natalists" emphasize the value of large families and warn of the societal dangers of a falling population. On the left, a countervailing "degrowth" movement has arisen, motivated by the fear of environmental catastrophe to encourage a slowing, stopping or unwinding of technological advancement and human settlement. Some even see civilization as a fundamental evil and advocate for voluntary human extinction.

Philosophers (at least a certain subset of them) study this question in the field of population ethics. Population ethicists analyze populations, or worlds, along two axes: size, how many people there are, and welfare, how good their lives are. They are specifically interested in comparing the overall quality of different worlds. Obviously this is a highly simplified model of real populations, which we might think of as more or less valuable depending on how virtuous and knowledgeable their citizens are, or how just their institutions are. But population ethicists assume worlds to be equal in all other respects, because while size is quite straightforward, welfare presents enough problems on its own.

[1] Parfit 1984, pg. 381.

Philosophers have come to no consensus on what exactly makes a person's life go well versus poorly, or whether this mysterious concept of "well-being"[2] can even precisely be defined. The first problem is easy to resolve. Population ethicists stay agnostic on the welfare debate, and make arguments which are compatible with any reasonable definition. The only requirement here is that interpersonal welfare comparisons are possible, and all this requires is that my life or yours is better than that of someone in a Siberian labor camp. The second problem is trickier, and I will have more to say about it later. For now it will be sufficient to observe that even with only an imprecise notion of well-being, it is clear some worlds are better than others. For instance, heaven (a huge world of extremely well-off people) is clearly better than hell (a huge world of extremely badly-off people). In this paper, I will use numbers to represent well-being, on the rough scale of 100 being a very good life and -100 being a very bad life, with 0 the dividing line between worthwhile and not-worthwhile lives.

The goal of population ethics is to formulate a "social welfare function" (SWF) which takes a world, and somehow combines each person's individual well-being into an aggregate value (V_W), which I will express in the unit of "points" as a "score" for that world, to allow it to be compared to other worlds (the higher the score, the more desirable a world). The simplest way to do so that takes everyone's interests into account is to attain global utility from a basic sum of each individual's well-being, so that everyone contributes exactly their personal well-being to the overall. This is the SWF known as "totalism," which I will defend.

But totalism is thought to suffer from a devastating objection: the Repugnant Conclusion, formulated by Parfit as he pondered the Awesome Question.

> *Repugnant Conclusion:* Compared with the existence of many people who would all have some very high quality of life, there is some much larger number of people whose existence would be better, even though these people would all have lives that were barely worth living.[3]

For example, imagine a world similar to heaven, but finite in size. Call this World A. Now, imagine a population a hundred times bigger, each with lives just three percent as good – though, critically, still good

[2] I will use the terms welfare, well-being, quality of life and utility interchangeably.

[3] Parfit 2016, pg. 110. This is a revised formulation, clearer than the original.

and worth living. Call this World Z. Since Z has three times as much well-being in it than A, by totalism it is three times as good. Yet this seems – well, repugnant.

The challenge for population ethicists is that avoiding the Repugnant Conclusion is not so easy as just rejecting totalism. In fact there are compelling arguments that repugnance is inescapable. In this paper, I will give one of these arguments and examine possible objections, with the hope of demonstrating why repugnance is so difficult to avoid. I will conclude that the best response to the Repugnant Conclusion is to accept it. Then, I will then contrast two conflicting intuitions which lie at the heart of population ethics, arguing in favor of the repugnant one. In the next section, I will introduce the problem of uncertainty which has not received the attention it merits in the field, and motivates my thesis. Finally, I will build up a modified totalist SWF that uses uncertainty to prevent repugnance from arising in most realistic scenarios, with the goal of softening the intuitive blow of accepting the Repugnant Conclusion.

§ ARRIVING AT THE REPUGNANT CONCLUSION

In this section I will offer a brief overview of the field of population ethics. First, I will provide a simple argument for the Repugnant Conclusion, and describe potential objections. Next, I will develop one of those objections into a popular non-repugnant SWF, and demonstrate its shortcomings. This discussion is intended to illustrate why progress in the field has been so difficult. Finally, I will briefly describe the status quo of population ethics, so that my argument may be placed in its proper context.

Here are three statements that, taken together, imply the Repugnant Conclusion:[4]

1. *The Benign Addition Principle:* If worlds *w* and *x* are so related that *w* would be the result of increasing the well-being of everyone in *x* by some amount and adding some new people with worthwhile lives, then *w* is better than *x* with respect to utility.

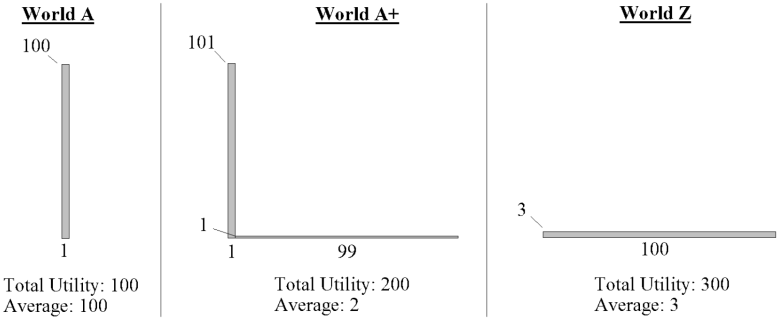
2. *Non-anti-egalitarianism:* If *w* and *x* have the same population, but *w* has a higher average utility, a higher total utility, and a

[4] Adapted from Huemer 2008, pgs. 2-4.

more equal distribution of utility than *x*, then *w* is better than *x* with respect to utility.

3. *Transitivity:* If *w* is better than *x* with respect to utility and *x* is better than *y* with respect to utility, then *w* is better than *y* with respect to utility.

This is Michael Huemer's "Benign Addition Proof", and though "proof" might be a bit strong, it's a quick, intuitive argument that well illustrates the challenge of avoiding repugnance. Here is a representation:[5]



In this figure, and all future ones, height represents well-being and width represents size. For consistency's sake, let's stipulate that each unit along the x-axis represents one billion people, so the size in A+ is 100 billion people, and the total utility is 200 billion. Now, we can attain A+ from A via Benign Addition – increasing the welfare of the billion people in A by 1 and adding 99 billion lives at welfare 1. From A+ to Z, we can see that total and average utility have increased, and complete equality has been achieved. From here a simple application of transitivity suffices to attain the result that Z is better than A, and repugnance is demonstrated, as any single Z-world being better than any single A-world is sufficient.

Though these steps are intuitive, all three can be objected to. You may have noticed that the move to A+ introduced glaring inequality. This can be mitigated by supposing that the two groups live on separate continents or even planets and are unaware of the other's existence, but it is true that inequality is a global feature of A+. If you see no benefit to creating new barely worthwhile lives, this may constitute a compelling objection. As such, in the next section I will

[5] Huemer 2008, pg. 4.

offer a defense of Benign Addition. In our second move from A+ to Z, we lose a certain high quality of life. A "perfectionist" objector might claim that A contains high-quality goods which cannot be exchanged for any amount of low-quality goods in A. The intuition here is that the Z-lives are "drab," free from pain and worry, but also almost entirely free from meaning and deep relationships. Parfit famously described these as full of "muzak and potatoes."[6] However, we may stipulate that the Z-lives are instead "roller coaster" lives, with the same peaks as in A, combined with a near-equal amount of deep suffering.

There's another way to object to Non-anti-egalitarianism, though. "Critical-level" theories claim that lives below a certain threshold should not count towards the score of the world. Usually the way this works is that each person's contribution to the aggregate is determined by subtracting that threshold from their well-being.[7] If you set the critical level high enough that it seems reasonable to prefer a large number of them to a smaller number of A-lives, then you've avoided repugnance.

However, critical-level theories have a problem of their own. The consequence of setting a positive threshold is that some worthwhile lives (Z-lives) decrease the score. This leads to what Gustaf Arrhenius termed the "Sadistic Conclusion":[8]

> *The Sadistic Conclusion*: When adding people without affecting the original people's welfare, it can be better to add people with negative welfare rather than positive welfare.

For example, let's assume that we have some base population and we can either add 1 billion lives at 1 or 10 million lives at -100. With a critical level of, say, 10, the first addition decreases the value of the base population by 9 billion while the second decreases it by 1.1 billion. So we should strongly prefer adding many extremely negative lives over worthwhile lives, which is even worse than repugnance. Out of the frying pan, into the fire.

Treating Z-lives as positive leads to repugnance, while treating them as negative leads to sadism. Critical-range theories try to sail through this Scylla and Charybdis by treating them neutrally. Specifically, all lives in the critical range between the critical level and 0 contribute nothing to the score. But surely a billion people at 10 is better than a billion people at 0? Population ethics is an objector's

6 Parfit 2016, pg. 118.
7 Totalism is a critical-level theory with a threshold of 0.
8 Arrhenius 2000, pg. 251.

paradise; it's very easy to formulate counterexamples to any proposed SWF, and impossible to formulate a SWF with no bullets to bite. Still, many people regard rejecting one of the premises of the Benign Addition argument as less painful than accepting its conclusion. That includes even transitivity, which was Parfit's preferred solution.[9] But this is not the only argument for repugnance.

The main sticking point for population ethicists has been the many "impossibility" arguments which derive an even more repugnant scenario (the "Very Repugnant Conclusion") from even more intuitive premises.[10] They are highly technical, so I won't try to reproduce one, but they are generally thought to succeed in producing a set of incompatible but seemingly necessary conditions for a satisfactory SWF, including anti-repugnance of some kind. Among these, anti-repugnance stands out. John Broome highlights the fact that the anti-repugnant intuition is very complex, compared to the simplicity of other conditions.[11] Mark Budolfson and Dean Spears argue convincingly that repugnance is a weird feature of aggregating over unbounded spaces, and should bear little practical force.[12] Then the easiest way to respond to these arguments, other than taking the nihilist way out and giving up, is to simply bite the bullet on repugnance. The goal of this paper is to soften that intuitive blow.

§ SIMPLICITY AND NEUTRALITY

In this section I will defend the Benign Addition Principle – the idea that adding worthwhile lives to a world increases its score. First, I will explain the "Simple View" (which entails Benign Addition) and the "Neutral View", two plausible but mutually exclusive intuitions. Then I will argue that we can embrace simplicity without many of the controversial views associated with it, and in so doing set up the main argument of this paper.

At the heart of repugnance are two conflicting intuitions. On the one hand, we have what Parfit terms the "Simple View":

[9] Parfit 2016, pgs. 114-115.
[10] See Arrhenius 2000, Arrhenius 2009, Budolfson and Spears (unpublished).
[11] Broome 2004, pgs. 57-59.
[12] Budolfson and Spears (unpublished), pg. 33.

> *The Simple View*: Anyone's existence is in itself good, and makes the world in one way better, if this person's life is good to live, or worth living.[13]

This seems quite plausible, but it's a one-way ticket to repugnance. If worthwhile lives always make the world better, then each life we add to Z increases its score, even if just by a tiny amount, and eventually it must eclipse A. Yet there is also what I will call the "Neutral View" proposed by Jan Narveson:

> *The Neutral View*: we are in favor of making people happy, but neutral about making happy people.[14]

This is also plausible, yet clearly inconsistent with the Simple View. If something makes the world better, then we should be in favor of it, all else being equal. The conflict seems to arise because each view considers the problem from a slightly different, well, viewpoint. When we think about each person's life from their own perspective, of course it seems that worthwhile lives are inherently valuable. But when we take more of a bird's-eye view, unless you're a natalist it seems strange to have a particular desire to enlarge the teeming mass of humanity.

One way of resolving the tension is to argue that the Simple View is too narrow. Sure, worthwhile lives matter, *if* they exist. This is simply the first part of the Neutral View. But why should we consider the existence of mere possible people? *They don't exist.* In fact, it doesn't even make sense to talk about their interests; those interests also don't exist![15] It's easy to get tangled in a metaphysical morass when discussing possible people, which is why I won't try to give a conclusive proof of the Simple View. But because the rest of my argument will have little force for those who reject it out of hand, I do feel compelled to offer a few points in favor of it.

First, the Neutral View turns out to be quite ambiguous when prodded. Simplicity can easily be formalized as totalism, but on the other hand, it's not clear how to represent neutrality as a SWF. We can interpret "in favor of making people happy" as endorsing something akin to Non-anti-egalitarianism, and "neutral about making happy people" as rejecting the Benign Addition Principle. But are we really neutral about making *any* happy people, or just barely happy

[13] Parfit 2016, pg 110.
[14] Narveson 1973, pg 80.
[15] Thanks to Prof. Samuel Asarnow for his helpful explication of these ideas.

people? If we are neutral about creating Z-lives but not neutral about creating A-lives, it looks like we've come back to critical-level totalism, with all its attendant problems. On the other hand, if we're hardcore neutralists, then we have some weird SWF that assigns all new positive lives a score of 0, and presumably all new negative lives some negative score. This leads to the unfortunate conclusion that having children always makes the world a worse place on expectation, because there's at least some chance of the child having a bad life, which I don't think is what most neutralists have in mind. There might be a different way of interpreting this view that threads the needle, but the lack of immediate clarity is not promising.

A related point is that the Neutral View doesn't hold up consistently. In small populations, it is much less plausible. A world of 50 happy people is surely better than a world of 10, better still 100. And all of these are preferable to 0 happy or unhappy lives for anyone who's not a radical degrowther. On the other hand, totalism can be upheld across the board, though we have seen that there are scenarios when it is tough to do so. Furthermore, the point where we stop caring about new happy lives seems to me precisely the point where we become completely unaware of their existence, and then of course we would rather see the lives around us improve rather than have more happy people created whose existence we will never notice. An intuition that bends itself to our self-interest in this way is not a reliable one from which to derive moral values.

Lastly, there are plenty of times we do in fact consider the interests of possible people. Consider the striving immigrant, who makes sacrifices to come to a wealthy Western country so that if and when she has kids, they will be better off. Or consider when governments craft forward-looking policy for the "future generations." If a nation's shrewd conservation policies help mitigate natural disasters 200 years from now, when the population, absent a dramatic technological breakthrough, will consist of entirely new people, it seems like that would be a good thing, even if no one alive now is affected. A neutralist might describe these as cases of making people happy rather than making happy people, as we are considering worlds where future people are better or worse off. The problem for this view, which Parfit terms the "Nonidentity Problem,"[16] is that the people in alternative futures are not guaranteed to be the same. When we consider how contingent our individual existences are (presumably a different sperm fertilizing one's mother's egg would have produced

[16] Parfit 1984, pgs. 351–380.

a different person, and very small changes in the past could've brought this about), it becomes clear that we are in fact considering cases of creating one group of people versus creating another (happier) group, with little or no overlap between the two. Preferring the latter is a rejection of neutrality.

An objection to the Simple View is that if possible people are allowed to enter into the equation, their sheer (potential) numbers quickly swamp the interests of those who currently exist. This is a counterargument often raised against "longtermist" philosophers, who support efforts to mitigate "existential risks" such as supervolcanoes or unaligned AI that threaten to wipe humanity off the map: in for a penny, in for a pound. If lowering x-risk is so important, why shouldn't we immediately divert all funding from education, healthcare and nutrition programs toward developing a global asteroid-defense system to avoid the fate of the dinosaurs? Without wading too deeply into that hot-button debate, I want to defend the Simple View's ability to avoid such absurd conclusions.

The Simple View is neutral in its own way – it does not care whether you make people happy or make happy people, just about how much well-being you create. And critically, in the real world it's often more efficient and less risky to make people happy than to make happy people. I can donate money to feed starving children right now, knowing that my generosity will have a substantial immediate effect. Conversely, if I, even as a well-off citizen of a prosperous society, were to have a child now, there would be a lot of dirty diapers in the near term and no guarantee that a happy life would eventually result from it. Even if the world where I have a child is likely better, it might not be the *best* I can bring into existence, which is why assigning high priority to natalism misses the mark. Practically speaking, I support a view that combines the attractive elements of both simplicity and neutrality, something like "we are in favor of making people happy, and in favor of making happy people once we've made everyone who currently exists happy."

A similar argument can shield simplicity from the dangerous implications of unrestricted longtermism. Even if there might be 10 quadrillion people sprawled across the universe in the year 5296, it is *extremely* uncertain what impact our actions now will have on them. Perhaps the best thing we can do now is improve the lot of just the next generation, which we can do with a high degree of confidence, so that they in turn will be prepared to look further into the future. The point is that the fact of one world being better than

another, while relevant, is not the only factor for making important decisions. In practice, these decisions often turn on uncertainty about which outcome will actually attain, rather than the best theoretically.

§ UNCERTAINTY

Now we have built up the foundations of population ethics, considered various proposed SWFs, and introduced the central problem of repugnance. We have offered a proof of that result, and dug into the intuitions behind it. In this section I will turn to my response to all of this. I will discuss the challenge posed by uncertainty, and then provide a thought experiment which should serve to illustrate my approach to dealing with it.

First, I will return to an important point I elided earlier, the challenge of precisely measuring well-being. If we are unable to do this, the entire exercise of population ethics is threatened. Though we have a general understanding of what constitutes a good versus a bad life, we are far from clear on a precise definition, or whether one is even possible. There are two explanations: either well-being is so complicated that we just don't know enough about it to formulate a definition, or well-being is inherently imprecise, and a life of quality 65 or -12 is a meaningless concept. Population ethics could survive if the latter were the case, as some worlds are clearly better than others and it would be nice to have an explanation why, but it would be much more convenient if the math we do was actually meaningful on some level. Still, even if a numerical scale of well-being is ultimately an idealization, idealizations can be very useful for high-level theories.

But idealization or not, if population ethics is to help us in the real world it must take into account this great uncertainty. And though uncertainty has received some treatment, in this area the literature is deficient. A notable result was Arrhenius' extension of his impossibility arguments over probabilistic outcomes. From similar premises, he managed to derive the even-more-than-very-repugnant "Risky Very Sadistic Conclusion," which presents a choice between A, or a lottery with a >99.99% chance of just the sufferers in Z- and a <0.01% chance of just the mediocre lives.[17] You could still eventually add enough mediocre lives that the expected sum of well-being by choosing Z- is greater than that of A. And Budolfson and Spears made a similar extension of their own impossibility result.[18] But the absurd

[17] Arrhenius and Stefánsson 2023, pg. 8.
[18] Budolfson and Spears unpublished, pgs. 15-17.

scenario of the Risky Very Sadistic Conclusion is even less realistic than the original repugnant and very repugnant results. I want to look at the practical implications of population ethics, and accordingly, I've tried to formulate an example which could plausibly resemble a decision humans will at some point have to make.

> *Space Colonization:* Sometime in the not-too-distant future, humanity builds a spaceship capable of transporting humans across the galaxy. In preparation for this monumental achievement, we sent rovers to seek out new planets to settle, and they've found and precisely analyzed two candidates. Planet A is quite similar to Earth but with much less land area. One billion people could eventually live there in great contentment, at welfare level 100. Conversely, Planet Z has much more land area, but is located at the frigid edge of the habitable range. One hundred billion people could eke out barely worthwhile lives there, at welfare level 1. The lives in A are 100 times better than those in Z, so the total expected well-being of both worlds is the same, but there is near-universal consensus that A should be chosen. Is this wrong?[19]

No. What kind of a lunatic would advocate for Z? As it turns out, in spite of the last 13 pages seemingly arguing for this, not me. Though I am a committed totalist, I see a subtle difference between *Space Colonization* and the Repugnant Conclusion. The key word here is "expected," because in this case and in all that could conceivably arise in the real world, we are basing our decisions off of projections and estimates rather than the certain numbers of thought experiments. And I see an important distinction here. Consider that the rover's estimate of well-being might be off by a small amount. Now we're much less sure *how* positive A's score will be, but it still will surely be positive. Z, on the other hand, could easily be quite negative if those barely worthwhile lives turned out actually to not be worth living. In the remainder of this paper, I will propose a method for incorporating uncertainty into totalism.

§ INCORPORATING UNCERTAINTY

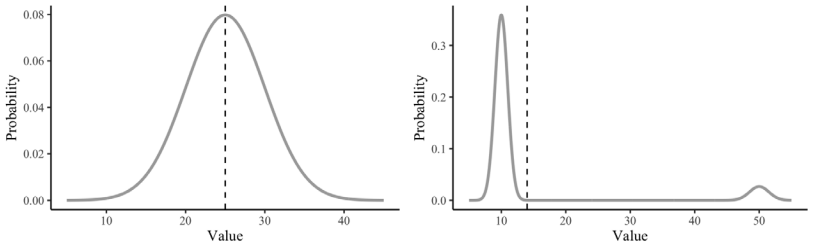
The remainder of this paper will focus on introducing and explaining my "risk-sensitive" approach to population ethics. Initially,

[19] Note: earlier, I analyzed a Z-world with lives of quality 3. I lowered that to 1 for this example to simplify the math. Both worlds are commonly regarded as repugnant.

I will summarize the math behind it and explain why I make some potentially controversial modeling choices. With this shiny new SWF in hand, I will apply it to *Space Colonization* and another counterexample to basic totalism,[20] but unfortunately conclude that my first pass does not sufficiently penalize repugnance. To address this, I will modify this first pass to create a new SWF that produces more (to my mind) satisfactory results.

However, while I feel very optimistic about my general risk-sensitive approach, there are many mathematical tools one could use to take uncertainty into account, and I do not claim that the SWF I present is the best possible.

The first piece of the puzzle is probability distributions (PDs). A PD allows us to assign varying degrees of likelihood to a range of possible outcomes, and represents them by mapping those outcomes to values on the x-axis and their likelihoods to heights on the y-axis. Since the total area under the curve is always 1, computing the area of different parts of a PD will give us the probability of an outcome in that range occurring. A common example is normal distributions, which can be used to model a surprisingly broad range of different things, from pinky lengths to standardized test scores to measured brightness of stars, and come in a distinctive bell-shaped curve.

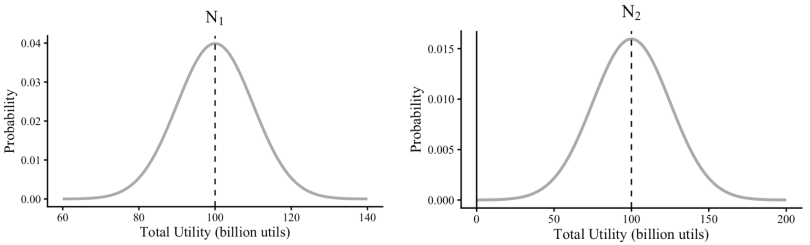


We will ask our rovers to represent uncertainty by reporting back a probability distribution (PD) of total utility rather than a single value. However, what we get may not be so neat as the normal distribution pictured above. But thankfully, for our purposes any PD will do. PDs contain a lot of information, but we will be focused on just two aspects. First is center, or more technically, the mean (represented by the dashed line). In the left figure, the mean is conveniently located at the peak right in the middle, but the geographic and numerical centers don't always coincide. This is because the mean is attained by

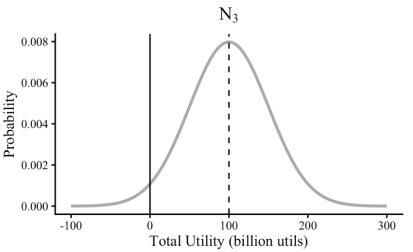
[20] Basic totalism can only handle worlds with no uncertainty, but there is a relatively straightforward way to extend it to probabilistic scenarios, which I will introduce later.

summing the product of each value along the x-axis with its probability along the y-axis. So in a bimodal distribution like the right figure, the mean may not be very close to the center of the picture. But it always represents the center of the data.

The other important aspect is the "spread" – essentially, the range of possible values. To illustrate why this matters, here are two more PDs.



Both are normal distributions with the same mean, but if you compare the values on each axis closely, you can see that the rover is less certain about its estimates of total well-being in N_2. The actual value of this planet is more likely to be worse off than predicted, but it's also more likely to be better off. Think about whether either seems preferable to you, then consider a third:



Again, the mean is the same, but this time the rover is even less confident – to the point that the world could actually be negative (just as likely, it could be extremely positive).

There are three possible routes one could take here. If N_1 seems better than N_2, which seems better than N_3, then you are *strongly risk-sensitive*. The fact that uncertainty exists at all makes a PD less preferable. If N_1 and N_2 seem similar, but both better than N_3, then you are *weakly risk-sensitive*. As long as the result is a positive outcome, uncertainty isn't a problem, but the possibility of creating a bad world is worth an extra effort to avoid. Finally, if the potential risks of all three PDs seem equally balanced with the potential rewards, then you

are *risk-indifferent*.[21] In the remainder of this paper, I will advocate for weak risk-aversion.

The theoretical case for risk-indifference is strong. If you see N_1, N_2, and N_3 as totally equivalent, then all you care about is the mean. And recall that the mean is essentially a weighted sum of the distribution. So risk-indifferent totalism is really the extension of basic totalism to probabilistic worlds, which is attractively simple and consistent if you accept basic totalism in certain worlds. And its proponents might push back against being termed indifferent to risk. The badness of the bad possibilities in N_3 does detract from the mean. It's just that there are proportional opportunities for much better worlds that balances it out.

The challenge for the risk-indifferent view is that most people care more about risk than this. Let's say I offer you a bet on a coin flip. If you call it correctly, you win $101, and if you don't, you lose $100. Do you take me up? Probably not, unless you really love to gamble. Yet if we modeled this bet as a PD, it would have a positive mean, and the risk-indifferent partisan would insist that it's irrational not to take the bet. You hesitate because, as Lara Buchak says, when we evaluate gambles we are "being sensitive to 'global' properties of gambles: [of] being sensitive not just to what happens in each particular outcome but to what the gamble looks like as a whole."[22] Buchak argues in the context of decision theory that it is in fact rational to take risk into account over and above its impact on the expected value, and I make a similar argument in the context of population ethics. The fact is that decisions about whether or not to take bets, or which planet should be colonized, do not take place in a vacuum.

To see why, let's raise the stakes of my coin-flip bet.

> *Well-To-Do Businessman:* A well-off businessman who's achieved a comfortable lifestyle is offered a bet on the outcome of a coin toss. If he guesses correctly, he will win $10 million, but if he is wrong, he will owe that much.

The businessman would recoil at the prospect. Sure, winning $10 million would be great. He could buy a bigger home, take fabulous trips, retire early. But being on the hook for $10 million would be catastrophic. He could lose everything, and still be shackled with a debt he could never pay off if he worked every day for the rest of his

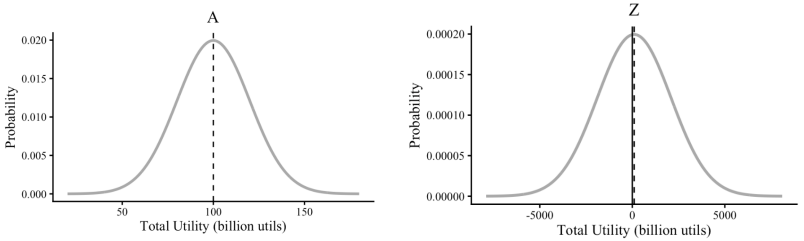
[21] Or perhaps, more bluntly, a *degenerate gambler*.
[22] Buchak 2017, 1.

life. Even though the monetary value appears the same, in reality the risk is much, much worse than the reward.

In *Space Colonization*, we are in the position of the businessman. He can walk away and get on with his comfortable life, and we can travel to A and be assured of a comfortable new colony. When presented with an alternative that could be much better but could also be much worse, he recoils, and rightfully so. Even if he were enticed with a slight gain in the expected value of the bet (+$10,001,000 vs -$9,999,000), he would be crazy to take it.

Since Z is necessarily much riskier than A, this scenario is analogous and can help explain why we resist repugnance. Recall that in the beginning we defined worlds along two axes: quantity, and quality. The critical thing to note is that while quantity is always positive (or 0), quality can be negative. Uncertainty about quality can produce uncertainty about the absolute goodness or badness of a world, while uncertainty about quantity merely affects the magnitude. And the score of Z is much more sensitive to fluctuations in quality than A, since its citizens live so much closer to the edge. All it would take is a slight error in the rover's extremely complex calculations to produce a net-negative world in Z, whereas A is all but guaranteed to be positive.

Let's take the scenario from *Space Colonization.* Suppose that the rover presents the choice between an A world with 1 billion people at welfare 100, and Z with 100 billion people at welfare 1. Additionally, the rover provides a confidence interval of ± 20 in its estimates of well-being, meaning that it is 95% confident that the people in A will live lives between 80 and 120, and equally sure that Z-lives will be between -19 and 21, and models uncertainty as normally distributed.[23] Then the PDs will look like this:



[23] It would also likely provide a confidence interval on its estimates of quantity. But since each additional person adds just 1 or 100 well-being, and each additional point of well-being can change the total value by billions or trillions, for simplicity's sake this can be ignored.

The rover is 99.9% confident that A's score will be at least 70 billion points, but only 54% sure that Z's will be greater than 0. The goal of our SWF is to properly penalize Z's possible negative outcomes.

To this end, *risk-sensitive totalism* generates the total value of a PD by taking the expected value of the PD, and multiplying it by an uncertainty modifier.

$$V_W = uE$$

This uncertainty modifier is derived by chopping off the negative range and computing the area under the distribution that remains. Specifically, we set u equal to the probability that any value X in the PD multiplied by E will be positive.

$$u = P(EX > 0)$$

Applying this formula to *Space Colonization*, every value in A's PD is positive, so $u = 1$ and the score remains at 100 billion points. But for Z, X is negative 46% of the time, so u is just 0.54, and the score is 54 billion. Conversely, in the original repugnant scenario there is no uncertainty and the scores for both remain the simple sums of well-being. Risk-sensitive totalism agrees with regular totalism in theory but deviates in practice.

An important aspect of this formulation is that u is always between 0 and 1. If u were greater than 1, then we would be rewarding rather than punishing uncertainty. And if u were negative, we would end up giving a positive on expectation scenario a negative score. On its face, this might not seem crazy. Perhaps we would prefer not to colonize any planet than to colonize one where we expect the people there to live lives of quality 0.01 on average, and possibly much worse. However, this also means that we could assign a negative world a positive score, which means we would accept the Sadistic Conclusion, which is a line I refuse to cross.

How should we, then, handle cases where E is negative? We must preserve the sign, but we can either continue to penalize uncertainty or reward it. Consider a "reverse repugnant" scenario where we are forced to choose between A- or a corresponding Z- world. This is somewhat analogous to T. M. Scanlon's well-known World Cup example, where several billion people are asked to give up fifteen minutes of watching the game to rescue someone in the transmitter

room of a TV station from excruciating electric shocks.[24] Most people agree that we should prefer the world where a bunch of people suffer small harms to the one where one person suffers a tremendous harm.[25] And rewarding uncertainty in negative cases is actually what our SWF currently does. While the PD of A- is always negative, Z- sometimes ends up being positive. Those positive outcomes decrease the value of u because we're now multiplying them with a negative E, which makes V_W less negative for more uncertain scenarios. This example illustrates how my SWF is risk-*sensitive* rather than risk-*averse*.

> Here's another case that illustrates the perils of risk-aversion:
>
> *Homeless man:* A homeless man living on the streets is offered a bet on the outcome of a coin toss. If he guesses correctly, he will win $10 million, but if he is wrong, he will owe that much.

Unlike the well-to-do businessman, the homeless man would jump at the chance to take this bet. He has nothing to lose, and everything to gain. Though the expected monetary value is the same as before, the agents' different treatments of risk makes these gambles completely different. And this same behavior can be observed across a wide variety of contexts. Consider a presidential election in which one candidate holds a clear lead over the other. The favorite will play it safe, avoiding confrontation and dreading any "October surprise" that could shake things up. Conversely, the underdog will throw anything and everything at the wall, hoping for just such a last-minute shakeup in the polls. Or consider a basketball game where one team has a double-digit lead in the final minutes. They will run down the clock with safe passes looking for easy layups while their opponents will aggressively hunt for steals and 3-pointers, trying to give themselves as many chances as possible to snatch victory from the jaws of defeat. We often give risk special consideration in our decision-making, but not always in the same way. Risk aversion can be just as much a vice as risk addiction.

We now have a working SWF, so let's evaluate its performance. We have already seen risk-sensitive totalism at work mitigating repugnance, and it has the additional virtue of mitigating a sort-of "utility monster" PD which could be contrived as a counterexample to basic totalism. Imagine that our rovers discover a planet M and report

[24] Scanlon 1998, pg. 235.

[25] Though I argue on the grounds of uncertainty rather than anti-aggregation – I don't share the intuition that *any number* of people in Z- is better than A-, no matter how uncertain.

a strange bimodal (two-peaked) PD giving a 99% chance of creating a world with mean V_W = -1,000, and a 1% chance of creating a world with mean V_W = 1,000,000,000. This PD has an expected value of 100,000 points. Alternatively, we could be guaranteed to create a world with 1000 people at welfare 99, producing 99,000 total points. Basic totalism tells us to leap at that 1% chance, but it seems quite counterintuitive to accept a 99% chance of creating a bad world. On the other hand, risk-sensitive totalism lets us pump the brakes. In the first scenario, the probability of creating a positive world is just 1%, so V_W = 0.01 * 100,000 = 1,000, much less than 99,000.

While this is quite a harsh penalty, you may have noticed a problem with its scoring of *Space Colonization*. With the numbers the way I originally laid them out, we prefer A to Z. But it is not difficult to adjust those numbers to produce the opposite result. All we need to do is double Z's size to 200 billion, and then its score will eclipse A's at 108 billion. And this still seems pretty repugnant. Fortunately, risk-sensitive totalism lends itself well to customization All we need to do is generalize the formula for deriving *u*:[26]

$$u = \frac{P(EX > 0)}{sP(EX \leq 0) + P(EX > 0)}$$

Here *s* represents a scalable risk sensitivity parameter (for any positive value). When *s* is 1, *u* is the same as before, because the denominator is always 1. As *s* increases, however, more weight is given to the negative possibilities. For instance, if you thought they should be given twice as much weight as the positives, then u_Z = 0.54/(2(0.46) + 0.54) = 0.37. Now Z would need to triple in size to catch up with A. You could push *s* higher and higher to confine repugnance to the most fanciful corners, but, as with everything in population ethics, this does come with a cost. The more sensitive to risk you become, the less sensitive you are to actual well-being (on expectation), and you'll become vulnerable to counterintuitive preferences for Z-like worlds with minimal uncertainty over worlds with much higher expected value and uncertainty that regular totalism strongly prefers.

However, in setting the value of *s* we are not confined to constants. As long as *s* is always positive, we can evade sadism. And

[26] Conservatively, neutral possibilities are categorized as negative rather than positive, but the effect on the math is negligible.

there is a compelling case to be made that *s* should scale with the mean. Currently, we penalize an outcome of -1 points equivalently to one of -100,000, simply by computing the probability of its occurrence. If *s* were to increase (and decrease) in magnitude with the mean, -100,000 points' stronger downward pull (assuming *E* is positive) will cause a corresponding rise in *s*. The challenge is how exactly to do this. If we just set $s = E$, since we're ultimately multiplying *u* by *E*, the mean terms will cancel. This would have the opposite effect of divorcing the score almost entirely from the expected value of the distribution, and this strange new SWF would not only be neutral about making happy people but also making people happy.[27]

The (somewhat arbitrary) function I landed on instead is the base-10 logarithm of the mean ($\log(E)$), which represents *E* as 10^x, and returns *x*. So if *E* is 10, it returns 2, if *E* is 10,000, it returns 4, and if *E* is 987654321, it returns 8.99. However, it's not as simple as just substituting *s* with $\log(E)$. Unfortunately, the base-10 log can return negative numbers for $E > 1$, because, for example, $0.1 = 10^{-1}$. To account for this, we must instead use the "log1p" variant, which takes the absolute value to account for negative values and then increases *E*'s magnitude by 1 so that when $E = 0$, $s = 0$, which is the desired behavior. My final proposed risk-sensitive totalism looks like this:

$$V_W = uE, \text{ where}$$

$$u = \frac{P(EX > 0)}{log(|E| + 1) * P(EX \leq 0) + P(EX > 0)}$$

Now we can return to *Space Colonization*. A retains its score of 100 billion points, because $P(EX \leq 0)$ is still 0%, so $u = 1$. For Z on the other hand, *u* works out to about 0.096. This produces a score of 9.6 billion points. That's substantially lower than 54 billion, and we can't just do the same trick we did last time and multiply Z's size by 11. Recall that the purpose of using a log function is to scale the penalty with the mean. Increasing the number of positive lives increases the expected value of the PD, and thus increases the modifier. With Z of population 1.1 trillion, we obtain *u* of roughly 0.089, and a score of 97.7 billion. It takes another thirty billion people for V_Z to eclipse V_A.

Is this a satisfactory result? I think so, with two caveats. First, if you just refuse to accept repugnance at all, no form of risk-sensitive

[27] As it solely represents the probability that a given world would have a score the same sign as the mean.

totalism is for you. In that case, I wish you good luck coming up with your own SWF that evades the impossibility arguments. Second, it's a little odd to draw a precise yet fairly arbitrary line where repugnance becomes acceptable (somewhere between 1.12 and 1.13 trillion people in Z). But I think this can be explained by admitting that even our rover's best estimates of their uncertainty are imprecise. When uncertainty disappears, so do arbitrary lines, as we're back to basic totalism. A practically useful SWF is not designed to be evaluated with such precision, it merely needs to be good enough in all cases.

With that said, let's evaluate whether the SWF is giving scores in the ballpark of reasonability. It expresses a preference for colonizing Z with 1.13 trillion people at (expected) welfare of 1, over A with 1 billion people at welfare 100. Strictly speaking, this is a repugnant result. But I think we would struggle to find a planet which could accommodate a trillion people, even at subsistence level. A better counterexample involves a much smaller A, likely a moon (though I imagine we would also struggle to find a moon which accommodates such a high quality of life). The SWF prefers 9.5 billion Z-lives to 10 million A-lives.[28] At this point, as a committed totalist I would argue that the potential value of Z now outweighs the risk. Usually, when someone offers you a choice between two bets, one with an expected monetary value 11.3 or 9.5 times that of the other, you should take that one (though of course there are businessman examples in which you shouldn't).

However, even if you're amenable to accepting repugnance in theoretical cases, you might desire to banish it entirely from real-world decisions. This could be accomplished by substituting another function for s, maybe by using the natural log rather than base 10. Now Z would not eclipse A until it reached a size of 3 trillion in the original *Space Colonization*, and 24 billion in the moon example in the previous paragraph. And this only scratches the surface of ways to customize my SWF. Perhaps, for instance, you favor the "strongly risk-sensitive" position I staked out earlier and want to penalize uncertainty itself, not just possible negative outcomes. To do so would require some mathematical wizardry, so I won't attempt such a modification here, but there's no reason it couldn't be done, and I won't even go so far as to say it shouldn't be done. My risk-sensitive totalism is only intended as a starting point, and I don't claim it should be the ending point, though its relative simplicity is an attractive feature.

[28] As E decreases, the denominator shrinks, which increases u, so there's less of a penalty.

The claim I do make is that uncertainty should be taken into account in any practical application of population ethics, in some form or another. This allows a strong but clean distinction between theoretical and practical population ethics, providing an easy response to the impossibility arguments, while potentially satisfactorily limiting the practical consequences of accepting repugnance. Furthermore, this view can go a long way, I believe, towards explaining why we resist the Repugnant Conclusion so strongly. When we read Parfit, we can't help but attempt to imagine the scenarios he presents in the real world, and in doing so unwittingly introduce significant uncertainty into the equation. Were we not worried about their proximity to negative territory, we would be better able to appreciate the value of the vast number of worthwhile lives offered by world Z.

REFERENCES

Arrhenius, Gustaf. "An Impossibility Theorem for Welfarist Axiologies." *Economics and Philosophy* 16 (2000): 247–66.

Arrhenius, Gustaf. "The Impossibility of a Satisfactory Population Ethics." In *Descriptive and Normative Approaches to Human Behavior*, 1–26. 2011.

Arrhenius, Gustaf, and H. Orri Stefánsson. "Population Ethics Under Risk." *Social Choice and Welfare* (2023).

Broome, John. *Weighing Lives*. Oxford: Oxford University Press, 2004.

Buchak, Lara. *Risk and Rationality*. Oxford: Oxford University Press, 2017.

Budolfson, Mark, and Dean Spears. "Why the Repugnant Conclusion is Inescapable." Unpublished manuscript.

Huemer, Michael. "In Defense of Repugnance." *Mind* 117, no. 468 (2008): 899–933.

Narveson, Jan. "Moral Problems of Population." *The Monist* 57, no. 1 (1973): 62–86.

Parfit, Derek. "Can We Avoid the Repugnant Conclusion?" *Theoria* 82, no. 2 (2016): 110–27.

Parfit, Derek. *Reasons and Persons*. Oxford: Oxford University Press, 1984.

Scanlon, T. M. "The Structure of Contractualism." In *What We Owe to Each Other*, 189–247. Cambridge, MA: Harvard University Press, 1998.